ISSN 1813 - 6796
ВІСНИК КНУТД №2 (108), 2017
Мехатронні системи. Енергоефективність та ресурсозбереження
Mechatronic Systems. Energy Efficiency & Resource Saving

УДК: 677.074

БІДЮК П.І., ДЕМКІВСЬКА Т.І., ДЕМКІВСЬКИЙ Є.О.
Київський національній університет технологій та дизайну

МЕТОДИКА ПОБУДОВИ РЕГРЕСІЙНИХ МОДЕЛЕЙ ЗА ЧАСОВИМИ РЯДАМИ

Мета. *Розробка методики побудови регресійних моделей часових рядів із застосуванням кореляційного аналізу даних та множини критеріїв адекватності моделей-кандидатів.*

Методика *ґрунтується на принципах системного аналізу даних та ієрархічного оцінювання структури і параметрів моделей.*

Результати. *Створено зручну для практичного використання методику побудови регресійних моделей часових рядів, яка забезпечує отримання адекватних моделей за умови наявності інформативних даних необхідної повноти.*

Наукова новизна. *Показана можливість застосування системного підходу до побудови регресійних моделей часових рядів і наведена послідовність виконання операцій при побудові моделей зазначеного типу.*

Практична значимість. *Запропонована методика забезпечує побудову моделей прийнятної адекватності за умови повноти та інформативності статистичних даних, представлених часовими рядами.*

Ключові слова. *Часовий ряд, побудова моделі, критерії адекватності, методика Дженкінса Бокса, прогнозування.*

Вступ. Для забезпечення умов прийняття високоякісних управлінських рішень необхідно удосконалювати існуючі та створювати нові методики моделювання на основі статистичних (експериментальних) даних.

Об’єкт та методи дослідження є статистичні (експериментальні) дані, які характеризують динаміку досліджуваних процесів. Методи дослідження: статистичний аналіз даних, методи оцінювання структури і параметрів математичних моделей і методи оцінювання прогнозів.

Постановка завдання. Для забезпечення належних умов для прийняття коректних управлінських рішень необхідно будувати адекватні математичні моделі процесів та об’єктів. Для створення методики моделювання часових рядів необхідно розв’язати такі задачі: скористатись принципами системного аналізу, забезпечити належну підготовку даних для моделювання, оцінити структуру і параметри моделей, а також сформувати критеріальну базу для коректного вибору кращих моделей із множини оцінених кандидатів.

Результати дослідження. Основи методики побудови моделей часових рядів запропоновані Боксом і Дженкінсом в роботі [1]. Ця методика використовується для побудови моделей фінансово-економічних, технологічних, соціальних та екологічних процесів, для яких характерно те, що для таких процесів не завжди можна отримати інформативні дані в достатньому об’ємі.

Модифікована методика побудови математичної моделі процесу, в основу якої покладено ідеї Бокса і Дженкінса, запропонована в [2]. Вона передбачає оцінювання структури і параметрів регресійних моделей часових рядів на основі кореляційного аналізу статистичних даних. Для оцінювання параметрів пропонується використовувати альтернативні методи, такі як метод найменших квадратів (МНК), метод максимальної

правдоподібності (ММП), нелінійні методи і метод Монте-Карло для марковських ланцюгів (МКМЛ) [3].

На основі вище наведеної методики в даній роботі запропоновано удосконалений алгоритм побудови регресійних моделей процесів, представлених часовими рядами. Відмінністю методики, що пропонується у даній роботі, є можливість врахування невизначеностей, які зустрічаються у процесі попередньої обробки даних і побудови математичних моделей.

Розглянемо етапи побудови моделі за часовими рядами.

1. Виконати аналіз процесу, для якого будується модель на підставі вимірів вхідних і вихідних змінних, представлених у вигляді часових рядів.

Аналіз процесу – це надзвичайно важливий етап, ігнорування якого призводить до неможливості побудови моделі високого ступеня адекватності. Він включає такі задачі:

– визначення кількості входів і виходів, тобто визначення розмірності процесу;

– встановлення логічних зв'язків між змінними та аналіз можливостей їх математичного опису (коректного об'єднання в одному математичному виразі);

– визначення кількості зовнішніх збурень та їх типу (детерміноване чи стохастичне, а також встановленні типу розподілу цих випадкових впливів);

– встановлення можливості декомпозиції процесу на окремі підпроцеси, які є простішими як з точку зору їх функціонування, так і з точки зору математичного опису; декомпозиція – це досить складний процес, який ґрунтується на спеціальних математичних методах;

– якщо процес має ієрархічну структуру (верхній та нижній рівень функціонування), то необхідно чітко розмежувати ці рівні, визначити функції кожного з них і встановити які типи зв'язків існують між ними; наприклад, технологічні процеси часто можна розмежувати на два та більше рівнів керування ними;

– використання знань із спеціальної літератури щодо особливостей функціонування процесу, відомих законів та закономірностей його протікання, виявлення існуючих моделей процесу та досвіду його теоретичного чи експериментального дослідження;

– за наявності розроблених моделей досліджуваного процесу необхідно встановити їх недоліки та переваги, а також визначити можливість подальшого використання (модифікації); аналіз і використання існуючих моделей надає можливість суттєво скоротити час та інші витрати на побудову та використання моделі.

2. Виконати попередню обробку даних.

Процес попередньої обробки (підготовки) експериментальних (статистичних) даних, як правило, включає такі операції:

– нормування та візуальну перевіркау даних і, при необхідності, їх корегування; нормування даних означає їх логарифмування або приведення до зручного діапазону їх зміни, наприклад, від 0 до 1; від –1 до +1; від +10 до –10 і т. ін.;

– корегування даних полягає у заповненні пропусків та зменшенні викидів (екстремальних значень), що виходять за основний діапазон значень змінних.

– формування перших або різниць вищих порядків, які необхідні для аналізу відповідних складових часового ряду; використання різниць надає можливість вилучити тренд із даних, тобто привести його до стаціонарної форми.

3. Оцінити структури моделей-кандидатів, для чого необхідно виконати такі операції: обчислити і виконати аналіз кореляційної матриці для часових рядів залежної і незалежних змінних із метою визначення змінних, які необхідно включити в модель; обчислити автокореляційну (АКФ) і часткову автокореляційну функцію (ЧАКФ), тобто значення коефіцієнтів кореляції в моменти часу *k-i, i* = 1, 2, …., для залежної змінної з метою вибору порядку авторегресії *p*.

4. Вибрати метод (методи) для оцінювання коефіцієнтів (параметрів) моделей-кандидатів і оцінити їхні параметри (МНК, ММП або інші).

5. Вибрати кращу (адекватну) модель з отриманої на попередньому етапі множини кандидатів, використовуючи для цієї мети множину відповідних статистичних критеріїв.

Критерії адекватності моделей. Діагностика моделей складається з таких кроків:

а) *Візуальне дослідження графіка похибок* моделі $e(k) = y(k) - \hat{y}(k)$, де $\hat{y}(k) -$ оцінка змінної, отримана за допомогою побудованої моделі. На графіку не повинно бути значних викидів та довгих інтервалів, на яких похибка приймає великі значення (тобто довгих інтервалів суттєвої неадекватності). У випадку застосування рекурсивних методів оцінювання найбільші похибки будуть у перехідному процесі, коли інформаційна матриця ще не містить достатньо інформації про процес.

б) *Похибки моделі не повинні бути корельовані* між собою. Для аналізу наявності кореляції між значеннями похибок необхідно обчислити АКФ та ЧАКФ для ряду $\{e(k)\}$ і за допомогою $Q-$статистики визначити ступінь корельованості (наприклад, $Q-$статистика вважається несуттєвою до рівня 10%).

Крім того, корельованість похибок визначають за допомогою статистики Дарбіна-Уотсона (DW), яка розраховується за формулою:

$$DW = 2 - 2\rho,$$

де $\rho = E[e(k)\,e(k-1)]/\sigma_e^2$ – коефіцієнт кореляції між сусідніми значеннями похибки; σ_e^2- дисперсія послідовності похибок $\{e(k)\}$. Таким чином, при повній відсутності кореляції між похибками $DW = 2$ – це ідеальне значення. Граничними значеннями для DW є 0 (при $\rho = 1$) та +4 (при $\rho = -1$).

Отримати формулу $DW = 2 - 2\rho$ можна досить просто. Автори цієї статистики (Durbin і Watson) запропонували скористатись для перевірки корельованості похибок моделі таким виразом:

$$DW = \frac{\sum_{k=2}^{N}\left[e(k) - e(k-1)\right]^2}{\sum_{k=1}^{N} e^2(k)} = \frac{\sum_{k=2}^{N}\left[e(k) - e(k-1)\right]\left[e(k) - e(k-1)\right]}{\sum_{k=1}^{N} e^2(k)},$$

тобто, DW можна, в деякій мірі, трактувати як коефіцієнт автокореляції для (перших різниць) приростів похибок.

Розкриваючи квадрат різниці в чисельнику, отримаємо:

$$DW = \frac{\sum_{k=2}^{N} e^2(k)}{\sum_{k=1}^{N} e^2(k)} + \frac{\sum_{k=2}^{N} e^2(k-1)}{\sum_{k=1}^{N} e^2(k)} - 2\,\frac{\sum_{k=2}^{N} e(k)\,e(k-1)}{\sum_{k=1}^{N} e^2(k)},$$

де $\frac{\sum_{k=2}^{N} e^2(k)}{\sum_{k=1}^{N} e^2(k)} \approx 1$; $\frac{\sum_{k=2}^{N} e^2(k-1)}{\sum_{k=1}^{N} e^2(k-1)} \approx 1$; а $\frac{\sum_{k=2}^{N} e(k)\,e(k-1)}{\sum_{k=1}^{N} e^2(k-1)} = \rho$.

Тому можна записати, що $DW = 2 - 2\rho$.

в) Для лінійної моделі 2-3 порядку *оцінки параметрів повинні збігатися до усталених значень* після 30-40 (не більше) ітерацій алгоритму оцінювання. Якщо кількість ітерацій набагато перевищує вказані числа, то це свідчить про те, що процес може бути нестаціонарним.

г) Перевірка значимості параметрів моделі. *Статистика Стьюдента або* $t-$*статистика* (випадкова величина, що має $t-$розподіл), яка використовується для визначення значимості оцінки кожного коефіцієнта в статистичному смислі, визначається за виразом:

$$t = \frac{\hat{a} - a^0}{SE_{\hat{a}}},$$

де $\hat{a}-$ оцінка коефіцієнта (параметра) моделі; a^0- нуль-гіпотеза (початкова гіпотеза) щодо цієї оцінки; $SE_{\hat{a}}-$ стандартна похибка оцінки. За нуль-гіпотезу щодо значимості оцінки можна висувати будь-яку: що коефіцієнт значимий, тобто, ($H_0 : a^0 \neq 0$) або незначимий ($H_0 : a^0 = 0$). Статистична теорія перевірки гіпотез пропонує висувати нуль-гіпотезу, яка є протилежною бажаному результату. В даному випадку бажаним результатом є значимість коефіцієнтів математичної моделі і адекватна модель. Таким чином, необхідно висувати таку нульову гіпотезу, що коефіцієнт незначимий. Це дає можливість коректно підійти до визначення значимості оцінок коефіцієнтів та дещо спростити розрахунки.

Для того щоб встановити, чи є оцінка коефіцієнта значимою, необхідно знати довжину вибірки даних N (потужність вибірки); число ступенів свободи $f = N - n$, де $n-$ число коефіцієнтів моделі, які оцінюються на основі ряду даних, і вибрати рівень значимості $\alpha = 1\%$ або $\alpha = 5\%$ або $\alpha = 10\%$ (для цих значень існують розраховані таблиці для критичних значень $t-$статистики). Фактично, рівень значимості означає ймовірність припуститись *помилки першого роду* при перевірці гіпотези. Згадаємо, що

$$\alpha = p\{X \in G/\omega \,|\, H_0\} = \int_{n-m(G/\omega)} L_{H_o}(X)\,dx,$$

де $X=[x_1,...,x_n]\in R^n-$ вся вибірка, яка розбивається на дві множини, що перетинаються: ω і G/ω ($\omega-$ область прийняття нуль-гіпотези); $G/\omega-$ критична область: якщо $X\in G/\omega$, то H_0 відхиляється; $L_{H_0}(X)-$ закон розподілу X. Помилка першого роду означає відхилення вірної гіпотези.

Користуючись значеннями $N,\ f$ і α, з таблиць для $t-$ розподілу знаходять критичне значення $t-$статистики, тобто $t_{кр}$. Для перевірки правильності висунутої гіпотези розраховане значення t порівнюють з критичним $t_{кр}$. Якщо

$$-t_{кр}<t<t_{кр}\quad \text{або}\quad |t|<|t_{кр}|,$$

то нуль-гіпотеза щодо незначимості коефіцієнта приймається (його можна не враховувати в регресії). Звідси випливає, що чим більшим є значення $t-$статистики для оцінки коефіцієнта, тим імовірніше, що цей коефіцієнт є значимим.

Загалом послідовність дій при перевірці значимості оцінок коефіцієнтів побудованої моделі можна сформулювати так:

– сформулювати нуль-гіпотезу щодо значимості коефіцієнта;

– обчислити значення $t-$ статистики для кожного коефіцієнта регресії (це робить кожний пакет для математичного моделювання);

– за допомогою значень $N,\ f$ і α знайти із таблиць для $t-$статистики її критичне значення;

– перевірити нуль-гіпотезу за наведеним вище простим правилом (аналіз виконання нерівності $-t_{кр}<t<t_{кр}$).

д) *Коефіцієнт множинної детермінації* R^2, який обчислюється так:

$$R^2=\frac{\text{var}(\hat{y})}{\text{var}(y)}=1-\frac{SSE}{SST},$$

де $\text{var}(\hat{y})-$ дисперсія залежної змінної, оціненої за допомогою побудованої моделі; $\text{var}(y)-$ дисперсія вимірів залежної змінної; $SSE=\sum_{k=1}^{N}[y(k)-\hat{y}(k)]^2-$ сума квадратів похибок (залишків) моделі (*sum of squared errors*); $SST=\sum_{k=1}^{N}[y(k)-\bar{y}]^2-$ загальна сума квадратів (*total sum of squares*); $\bar{y}-$ середнє значення; $SST=SSE+SSR$, де $SSR=\sum_{k=1}^{N}[\hat{y}(k)-\bar{y}]^2-$ загальна сума квадратів для регресії (*sum of squares for regression*).

Очевидно, що найкращим значенням є $R^2=1$, тобто, коли дисперсії вимірів змінної, та цієї ж змінної, оціненої за рівнянням, збігаються. Цей параметр можна трактувати, також, як міру інформативності моделі, якщо вибрати за міру інформативності дисперсію. Таким

чином, R^2 показує рівень інформативності моделі по відношенню до інформативності вибірки даних, за допомогою якої вона була оцінена.

е) *Сума квадратів похибок* для вибраної моделі повинна бути мінімальною, тобто

$$\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2 \to \min_{\hat{\theta}}$$

порівняно з усіма іншими моделями.

є) Для оцінки адекватності моделі також використовують *інформаційний критерій Акайке*

$$AIC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n$$

та *критерій Байєса-Шварца*

$$BSC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + n \ln(N),$$

де $n = p + q + 1-$ число параметрів моделі, які оцінюються за допомогою статистичних даних (p - число параметрів авторегресійної частини моделі; q - число параметрів ковзного середнього; 1 з'являється тоді, коли оцінюється зміщення (або *перетин*), тобто a_0).

Критерії Акайке і Байєса-Шварца містять в правій частині суму квадратів похибок, а тому за цими критеріями вибирають ту модель, для якої критерії приймають найменші значення. Введення нового регресора приводить до збільшення критерію (при цьому збільшується n), але, разом з тим, зменшується сума квадратів похибок і критерій в цілому зменшується. Якщо регресор не покращує модель, то критерій збільшується. Необхідно також зазначити, що асимптотичні властивості для довгих виборок кращі у критерія Байєса-Шварца, тобто, його рекомендують застосовувати при відносно великих значеннях N ($N > 100$).

ж) Окрім згаданих параметрів, для визначення адекватності моделі в цілому використовують $F-$*статистику Фішера*, яка пропорціональна відношенню:

$$F \sim \frac{R^2}{1 - R^2},$$

а для множинної (багатофакторної) регресії вона визначається за виразом

$$F = \frac{R^2}{1 - R^2} \cdot \frac{(N - p - 1)}{p},$$

де, як і раніше, $N-$ число значень ряду; $p-$ число параметрів моделі без врахування перетину (константи).

Таким чином, якщо $R^2 \to 1$, то $F \to \infty$. Порядок застосування $F-$статистики такий же, як і $t-$статистики. Нуль-гіпотезою є в даному випадку припущення про те, що модель неадекватна в цілому, тобто,

$$H_0 : a_1 = a_2 = ... = a_p = 0$$

проти альтернативної гіпотези

H_1: *хоча б одне значення* a_i *відмінне від нуля в статистичному смислі*.

Значення $F_{крит}$ знаходять із таблиць для $F-$ розподілу. Послідовність застосування цієї статистики відповідає загальній процедурі перевірки статистичних гіпотез.

Коректне застосування методики Бокса-Дженкінса забезпечує побудову адекватної математичної моделі процесу, якщо експериментальні дані відповідають *вимогам представництва* та *інформативності*. Перша вимога означає, що вибірка даних повинна охоплювати досить довгий проміжок часу, щоб повністю відображати поведінку того режиму фукціонування процесу, для яких будується модель. Вимога *інформативності* означає, що вибірка повинна містити в собі об'єм інформації, достатній для оцінювання коефіцієнтів моделі. Наприклад, якщо моделюється процес другого порядку, то вибірка повинна забезпечувати коректне обчислення першої та другої похідної. Іноді формально інформативність оцінюють за допомогою величини дисперсії процесу, а також за кількістю гармонічних складових, які містяться в процесі. Чим більше гармонік містить вибірка, тим вищою є її інформативність.

Умову інформативності даних пов'язують з *умовою достатнього збудження* процесу. Достатнє збудження означає, що вхідний сигнал повинен охоплювати всю смугу частот, які може пропускати на вихід процес (об'єкт). Тобто вхідний сигнал повинен охоплювати всю амплітудно-частотну характеристику процесу. Ця вимога залишається однаковою для процесів будь-якої природи. Приклади побудови математичних моделей за часовими рядами будуть подані у наступній статті.

Висновки.

1. Розроблена методика побудови регресійних моделей за даними у формі часових рядів. Вона складається з п'яти етапів: аналіз досліджуваного об'єкта, попередня обробка даних, оцінювання структури моделі, оцінювання параметрів (коефіцієнтів) моделі і діагностика побудованих моделей кандидатів з метою вибору кращої альтернативи.

2. Попередня обробка даних – це необхідний етап моделювання, який забезпечує приведення даних до форми, зручної для виконання подальших обчислень. Цей етап забезпечує заповнення пропусків даних, обробку екстремальних значень, нормування та фільтрацію вимірів.

3. Оцінювання структури моделі включає аналіз кореляційної матриці з метою вибору регресорів для включення у модель, визначення порядку авторегресії за допомогою автокореляційної та часткової автокореляційної функцій, а також виявлення не лінійності і ідентифікацію випадкових складових у вимірах, тобто визначення типу розподілу зовнішніх збурень.

4. Наведені статистичні критерії якості для оцінювання адекватності моделей-кандидатів, використання яких забезпечує вибір кращої моделі для подальшого обчислення оцінок прогнозів.

5. Загалом коректне використання запропонованої методики забезпечує отримання моделей високого ступеня адекватності за умови достатньої інформативності та повноти статистичних (експериментальних) даних.

6. При виконанні майбутніх досліджень доцільно модифікувати методику на побудови моделей нелінійних процесів і застосувати ймовірнісні моделі для отриманні додаткових альтернатив при прогнозуванні.

Список використаної літератури

1. Box, George; Jenkins, Gwilym Time series analysis: forecasting and control, rev. ed. // Oakland, California: Holden-Day. — 1976. – 712 p.

2. Бідюк П.І., Романенко В.Д., Тимощук О.Л. Аналіз часових рядів: Навчальний посібник. – Київ.: НТУУ КПІ, 2013. – 599 с.

3. Tsay R.S. Analysis of financial time series. – Hoboken, New Jersey: John Wiley & Sons Inc., 2010. – 715 p.

References

1. Box, George; Jenkins, Gwilym Time series analysis: forecasting and control, rev. ed. // Oakland, California: Holden-Day. — 1976. – 712 p.

2. Bidyuk P.I., Romanenko V.D., Timoshchuk O.L. Analiz chasovikh ryadiv: Navchal'niy posibnik. – Kiïv.: NTUU KPI, 2013. – 599 s.

3. Tsay R.S. Analysis of financial time series. – Hoboken, New Jersey: John Wiley & Sons Inc., 2010. – 715 p.

МЕТОДИКА ПОСТРОЕНИЯ РЕГРЕССИОННЫХ МОДЕЛЕЙ ПО ЧАСОВЫМ РЯДАМ

БИДЮК П.И., ДЕМКОВСКАЯ Т.И., ДЕМКОВСЬКИЙ Е.А.
Киевский национальный университет технологий и дизайна

Цель. *Разработка методики построения регрессионных моделей временных рядов с применением корреляционного анализа данных и множества критериев адекватности моделей-кандидатов.*

Методика *основывается на принципах системного анализа данных и иерархического оценивания структуры и параметров моделей.*

Результаты. *Создана удобная для практического использования методику построения регрессионных моделей временных рядов, которая обеспечивает получение адекватных моделей при наличии информативных данных необходимой полноты.*

Научная новизна. *Показана возможность применения системного подхода к построению регрессионных моделей временных рядов и приведена последовательность выполнения операций при построении моделей указанного типа.*

Практическая значимость. *Предложенная методика обеспечивает построение моделей приемлемой адекватности при полноты и информативности статистических данных, представленных временными рядами.*

Ключевые слова. *Часовой ряд, построение модели, критерии адекватности, методика Дженкинса-Бокса, прогнозирование.*

METODS OF DEVELOPMENT REGRESSIVE MODELS OF TIME SERIES

BIDYUK P.I., DEMKIVSKA T.I., DEMKIVKIY E.A.
Kyiv National University of Technology and Design

Objective. *Development of construction methods of time series regression models using correlation data analysis and adequacy criteria set of candidate models.*

Methods *is based on principals of system data analysis and hierarchical evaluation of structures and parameters of the models.*

Results. *A convenient for practical use method of construction regression time series is developed that provides obtaining adequate models subject to the availability of necessary informative data which is enough complete.*

Scientific novelty. *The possibility of a systematic approach to developing regression models of time series is shown and the sequence of operations under the construction of stated models is described.*

Practical significance*. The method provides development of models with suitable adequacy on the base of complete and informative statistical data presented by time series.*

Keywords. *Time series, building of a model, the adequacy criteria, The Box-Jenkins method, prediction.*